图书情报工作
第65卷 第11期 2021年6月

融合 PageRank 与评论情感倾向的在线健康社区用户影响力研究*

■ 董伟 陶金虎
天津大学教育学院 天津 300350

摘 要：[目的/意义] 在线健康社区中对高影响力用户的有效识别，有助于健康信息需求者发现有价值的健康信息，对于降低健康信息查找成本和提高健康行为决策的有效性具有重要意义。[方法/过程] 从用户交互性和评论情感倾向出发，利用 PageRank 和 SVM 等算法构建出在线健康社区用户影响力的测量方法，并以医享网为实验对象，从发布内容使用价值的视角，进一步计算了该社区中用户的综合影响力，并对案例用户进行分析。[结果/结论] 分析结果表明该算法具有一定的合理性，能够对 PageRank 算法的影响力计算结果进行优化；同时，利用 TF-IDF 和互信息算法揭示了高综合影响力用户发布的信息内容与社区其他用户群体内容主题基本一致，该类用户对社区的主题方向起到一定的引导作用。因此，通过本研究所构建的方法可以有效识别高影响力的用户，有助于健康信息需求者及时准确的发现所需信息，提高健康信息的使用效果，从而丰富在线健康社区用户信息行为的理论和实践研究。

关键词：PageRank 情感倾向 在线健康社区 用户影响力
分类号：G252
DOI：10.13266/j.issn.0252-3116.2021.11.002

1 引言

"互联网+医疗"发展战略是顺应时代的产物，也是向智能医疗转变的必经之路。用户不仅可以在线预约挂号、查阅资料，还能够得到意向领域专家的解答或者病友的经验传授与讨论，缩短传统医疗中寻根问药的时间成本，大大提高了用户的参与感与治疗效率。据医疗相关数据显示，2018年全国超过99万家卫生机构总会诊人次达到33.8亿[1]，2019年在线咨询总量达5.6亿次，未来将持续保持上升趋势[2]。同时，《"互联网+医疗健康"发展的意见》[3]也鼓励在线健康社区运用互联网的相关技术加快实现资源互通、信息共享与远程医疗等服务，不断健全互联网+医疗的一体化服务体系，加强医院、医生与患者间的有效沟通。

目前，国内健康问题讨论规模较大的在线社区以医享网、39健康论坛与好大夫在线等网站为主，这些在线社区用户多，知识传播速度快，产生了大量的信息与数据，为健康信息需求用户提供了有价值的健康信息。在线社区中存在一些活跃程度较高的用户，他们能够吸引到其他用户的关注和互动，从而在一定程度上影响其他用户的信息行为和健康决策，对于整个在线社区的信息传播具有较强的导向作用。然而，用户的活跃程度与其所发布的信息的使用价值间并非存在直接关系，如一些用户具有较强的交互影响力，在社区活跃程度较高，求助和抒发情感等行为频繁，所发布的信息也受到较多关注，但其他用户对其评价不高，在一定程度上反映了其信息的使用价值有限；还有一些用户尽管交互活跃程度不高，但其所发布的信息受到的积极评价较多，其所发布的信息具有较好的应用价值。因此，从信息使用价值角度出发，如何结合用户活跃性和交互情感倾向性识别来判断在线健康社区用户的综合影响力，对于帮助用户便捷、有效地利用健康信息，做出客观的健康行为决策等方面具有重要意义。本研究拟在融合用户交互活跃性和评论情感倾向的基础上，探索性地构建在线健康社区用户综合影响力的测量算法，并在相应的在线健康社区中进行实验和结果

* 本文系国家社会科学基金青年项目"在线健康社区用户交互行为及其对用户健康效用影响研究"（项目编号：16CTQ029）研究成果之一。
作者简介：董伟（ORCID：0000-0002-7632-2386），副教授，博士；陶金虎（ORCID：0000-0003-4316-0795），博士研究生，通讯作者，E-mail：tcarry@tju.edu.cn。
收稿日期：2020-11-23 修回日期：2021-02-05 本文起止页码：14-23 本文责任编辑：杜杏叶

董伟，陶金虎. 融合 PageRank 与评论情感倾向的在线健康社区用户影响力研究[J]. 图书情报工作,2021,65(11):14 - 23.

分析,以期为有效挖掘在线健康社区中有影响力的用户和有价值的健康信息提供一定的方法和参考。

2 相关研究

用户影响力的分析与测量是在线社交媒体和在线社区相关研究领域中,学者所关注的重要研究方向之一。目前关于用户影响力的相关研究主要采用特征值统计分析方法、社会网络分析方法以及 PageRank 方法等。

特征值统计分析方法主要通过统计能够反映在线社区用户活跃特征的相关特征值,并进行一定的指标和权重的设定,从而计算用户的影响力。如王佳敏等[4]在分析用户影响力时,主要统计了影响力指标和活跃度两个指标,其中影响力指标包括粉丝数、被转发数、被评论数、是否认证 4 个特征值,活跃度指标包括微博数和关注人数两个特征值。赵发珍等[5]利用博客的引用数量、回复数量、网页内外链接数等特征值进行用户影响力的建模。董伟等[6]也通过获取和分析在线社区中用户的留存时间、发帖量、粉丝数等反映个人和交互维度的相关特征值,对活跃用户进行了识别,并对其在社区中的影响力进行了分析。

社会网络分析方法主要通过关系网络结构中的属性值来计算各个网络节点在网络中的重要性,如网络密度、点度中心性、中介中心性、接近中心性及等。陈远等[7]通过分析社会网络的中心度、结构洞等指标来挖掘在线社区中用户的影响力。谢英香等[8]则通过对社会网络分析法中的中心度的分析,利用 MDS 等方法,分析了虚拟社区中的用户的影响力,并进一步揭示该社区存在意见领袖现象。S. Jonnalagadda[9]等则综合分析了点度中心性、点度中介性、以及点度紧密性等反映中心的指标,从而发现了医学在线社区中具有较大影响的意见领袖。

PageRank 算法认为,社交网络中用户间的点赞、转发与评论等互动关系与网页之间的链接指向非常类似,因此网页间链接结构的分析方法也可以用于社交网络用户之间转发、评论等互动关系的分析[10]。PageRank 算法也被越来越多学者应用于在线社区用户影响力的分析和测量等方面。如刘玲等[11]、张俊豪等[12]在 PageRank 算法的基础上融入了用户行为中转发率、评论率、微博数量、时间间隔等指标,对微博社区中信息传播核心贡献者和高影响力用户进行了探索;X. Song 等[13]则通过将用户提供的信息新颖性与 PageRank 相结合,提出了综合影响力算法。苑丽玲等[14]在 PageRank 算法基础上考虑了加权社会网络相关因素,对 PageRank 算法进行了改进,从而对用户的影响力进行了探索。肖宇等[15]在 PageRank 的基础上进一步考虑了用户之间的互动程度以及用户共享意愿,从而提出了用于计算用户影响力的 Weibo-Rank 算法。

综上所述,当前关于用户影响力分析的研究主要集中于对用户互动指标、交互网络结构属性的分析,但多数研究都主要从单一的视角对在线社区用户的影响力进行评价和分析,这会在一定程度上降低用户影响力测量的有效性。特征值统计分析方法与社会网络分析方法虽能在不同程度上衡量社区用户影响力,但前者过于依赖特征得分,忽略了真实交互影响力,后者多聚焦小型网络,更多侧重于直接关系的测量。而 PageRank 算法既支持计算交互影响力,也能够融入更多特征得分,具有较好的融合性,能够较为客观全面的反映出用户综合影响力。在线健康社区中,用户之间的交互信息情感倾向可以有效的判断社区中信息是否具有良好的利用价值,但当前多数研究中忽略了该类主观因素。因此,有必要对用户交互行为和评论情感倾向进行结合,进一步完善和发展在线健康社区用户影响力计算和评价方法。故本研究从交互性和情感倾向融合的视角出发,先使用 PageRank 算法对在线健康社区中所有用户的交互影响力进行排序,然后通过判断和寻找最优机器学习情感分类模型,识别用户评论情感倾向性,进而融合交互影响力和情感倾向计算和识别出用户的综合影响力。

3 研究设计

3.1 研究思路

本研究的思路主要包括四个步骤,首先是利用数据爬虫对在线社区相关信息进行爬取,对数据进行预处理,并将最终可用的数据存入数据库,包括用户和评论信息两个方面。其次,是对用户的综合影响力进行计算,综合影响力主要包括三项子算法:①利用 PageRank 算法对用户的交互影响力进行计算;②通过选择最优情感分类模型,对评论信息进行情感归类与分析,并进一步对评论信息情感倾向值进行计算;③融合上述两内容的结果按照特定公式进行融合,并通过案例分析进行对比。再次,利用 TF-IDF 与互信息算法进一步探究高综合影响力用户所生产的信息内容与社区其他用户群体内容主题方向的关系,并通过可视化的方法进行比较分析。最后,对本研究的研究过程和方法进行总结,并提出相应的研究展望。如图 1 所示:

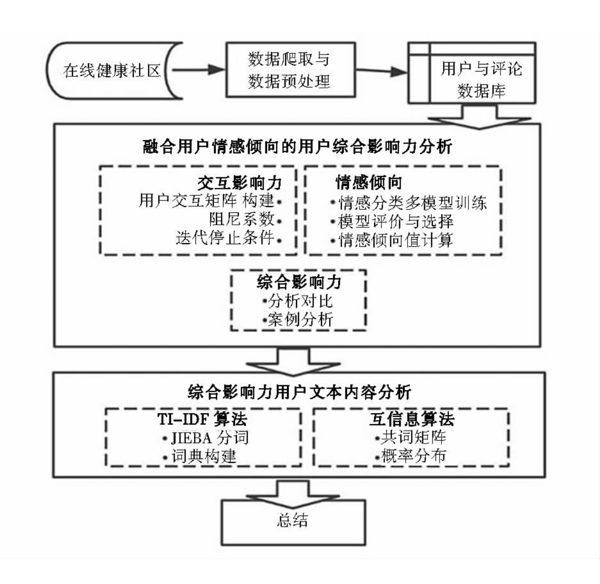


图1　研究思路

3.2　数据获取与预处理

本研究以健康社区中的用户所发布的信息及其评论信息为分析对象，使用 Python 语言构建多线程爬虫工具，以 Cookie 参数与报头信息作为用户与浏览器表征工具，通过解析 DOM 树获得该社区中用户交流之间的相关内容，包括用户昵称、发帖内容与相应的回帖信息。

此外，进一步对相关数据进行预处理，如分词处理、用户编码映射表构建、用户评论映射表构建、用户评论者映射表构建、异常用户处理等。本研究拟以医享网社区的用户生成内容为例，并收集相应数据进行相关实验和分析。

3.3　分析过程与技术

传统 PageRank 算法中，较多考虑的是网站或者用户之间的交互关系与权重，并不对其本身质量进行分析，故本研究结合用户交互关系与用户评论等信息内容进行分析，一方面发掘潜在网络用户影响力排名；另一方面对用户情感倾向进行识别，并融合两者进行综合性探究。

3.3.1　用户交互影响力的计算

本研究抽取了用户与评论用户的映射关系，对发帖人与评论人的多元共现关系进行了梳理，并将具体交互网络转换为交互矩阵，使用 PageRank 算法得到交互网络中交互影响力较高的用户，具体算法如下：

基本 PageRank 算法思想如公式 1 所示，其中 O 表示其他指向 A 节点的节点，$PR(O)$ 表示其他指向 A 节点对应的节点 PR 值，$L(O)$ 表示其他指向 A 节点的节点出链数，PR′为对应节点的下一次迭代 PR 值，m 表示模型收敛时的迭代次数，N 表示用户节点的总个数，各用户节点初始 PR 值为 1/N，最终 PR 值为这些用户的交互影响力得分。

$$PR' = \sum_{i=n}^{m} \frac{PR(O)}{L(O)} \qquad \text{公式(1)}$$ [16]

为方便计算一般使用公式(2)(等价转换于公式(1))的形式进行计算，其中 M 为本研究中用户交互网络形成的转移矩阵，具体如公式(3)所示，$M(u_i,u_j)$ 表示用户 j 出链到用户 i，即用户互动情况，PR 为 PR' 上一次迭代结果。

$$PR' = M * PR \qquad \text{公式(2)}$$

$$M = \begin{bmatrix} M(u_1,u_1) & M(u_1,u_2) & \cdots & M(u_1,u_N) \\ M(u_2,u_1) & \ddots & \cdots & M(u_1,u_2) \\ \vdots & \vdots & M(u_i,u_j) & \vdots \\ M(u_N,u_1) & \cdots & \cdots & M(u_N,u_N) \end{bmatrix}$$

公式(3)

然而，上述计算方式对于某些入链自身节点的 PR 值解释无力，并造成节点的 PR 值产生偏移和错误，入链自身节点的 PR 值最终为 1，而其他节点 PR 值为 0。为解决这一问题，引入公式(4)进行修正，其中，β 为阻尼系数，取值 0.85，主要用于解决陷阱与孤立点问题。

$$PR' = \beta * M * PR + \begin{bmatrix} (1-\beta)/N \\ (1-\beta)/N \\ \vdots \\ (1-\beta)/N \end{bmatrix}$$

公式(4)[16]

设置迭代停止条件为下一次 PR' 值与上一次 PR' 值相等，且 $\sum_{p=1}^{N} PR_p' = 1$。该算法能有效发掘交互网络中的关键人物，并对这些用户能被赋予较高的 PR 值，从而发现交互影响力较高的用户。

3.3.2　用户评论文本情感倾向识别的计算

在线健康社区用户生成内容的评论存在着明显的情感倾向，而这类倾向可以作为用户影响力及生产内容质量评价的重要指标之一[17]。本研究通过对文本进行大量特征抽取基础上，借助有监督机器学习模型进行情感倾向识别，包括随机森林算法、Logistic 算法、SGD 算法、SVM 算法、朴素贝叶斯算法。性能评价指标包括准确度和 F1 值，其中 F1 值与召回率与精确率均有关，一般被认为是评价模型优劣的综合性指标，计算方式见公式(5)与公式(8)，TP 指正类被预测为正的数量，TN 指负类被预测为负的数量，FP 指负类被预测为正的数量，FN 指正类被预测为负的数量。情感分类类

别主要涉及三方面:①表示支持,标记为 1;②商讨、讨论表示中立,标记为 0;③反对表示否定,标记为 2。

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$ 公式(5)[18]

$$Precision = TP/(TP + FP)$$ 公式(6)[18]

$$Recall = TP/(TP + FN)$$ 公式(7)[18]

$$F1 = (2 * Precision * Recall)/(Precision + Recall)$$ 公式(8)[18]

基于上述评价指标选择合适的模型进行预测,并对结果进行梳理,所公式(9)所示,该计算思想能够克服不同比例和不同数量级上数据带来的干扰。其中,AV 代表情感倾向值,U_p 表示所有用户中某一位用户,w 属于 0 或者 1 类,即非否定类,len(w)用于衡量具体类别个数,r 表示这位用户收到的评论情感类别。为进一步降低数量级关系带来的干扰,统一将每位用户的情感倾向值放入列表中,并通过标准化函数进行归一化。

$$AV = \sum_{w=0}^{1}\frac{U_p(len(w))}{U_p(len(r))}U_p(len(w))[*2\ if\ w == 1],r \in [0,1,2]$$ 公式(9)

3.3.3 融合两种算法的综合影响力的分析

本研究对上述 PageRank 算法和情感倾向识别结果进行融合,以期探索从交互性和评论的情感性两个方面对健康社区用户影响力进行综合性的评价。在 PageRank 算法基础上融合情感倾向值,即将用户的情感倾向值作为相应用户的权重,与交互影响力进行融合,形成用户新的综合影响力值,见公式(10),其中 p 表示 n 个用户中的某一位,UR 代表综合影响力,PR_p' 代表交互影响力,AV_p 代表情感倾向值。

$$UR_p = PR_p' * AV_p, p \in [1,2,3,\cdots,n]$$ 公式(10)

3.3.4 综合影响力用户文本内容的分析

为进一步探究高综合影响力用户对健康社区主题方向的影响,即这些具有高综合影响力的用户信息文本是否在一定程度上代表或影响了社区的内容主题方向,本研究将进一步使用 TF-IDF 和互信息算法构建不同用户群体生成内容的共词矩阵进行分析和对比。首先使用 TF-IDF 分别计算高综合影响力用户和社区所有用户生成内容的高频词,然后借助互信息算法抽取高频词最为相关的若干词条,从而分别形成高综合影响力和社区其他用户群体内容的共词矩阵网络,并对其做进一步的比较,以探索高综合影响力发布内容与社区其他用户发布内容的关系。

(1)TF-IDF 计算。TF-IDF 是一种加权算法,它的优点在于可以过滤掉文本中常见但没有实际意义的词语,同时保留真正影响文本的词语,因此 TF-IDF 相较于普通的词频统计更加准确和客观,具体算法如下:

$$TF\text{-}IDF = \frac{N_{i,j}}{\sum_1^k N_{k,j}} * \log\frac{D}{D_i + 1}$$ 公式(11)[19]

其中,$N_{i,j}$表示用户生成关键词 i 在文档 j 中的出现频次,$\sum_1^k N_{k,j}$表示 k 个关键词对应的文章总词数,即前半部分计算称为 TF,表示关键词 i 在文档 j 中出现的频率。后半部分 D 表示语库中的文档总数,D_i 表示 D 篇文档中包含关键词 i 的文档数量,同时为避免所有文档都不包含该词,故分母加 1。

(2)互信息。互信息主要指的是知道一个词条,而对另外一条词条的不确定性减少的程度。具体而言,需要我们先使用 JIEBA 对用户生成内容进行分词,然后遍历带有高频性质的词条与其他分词间相互依赖性的度量,并在此基础上形成高频词－互信息网络,以此进行可视化和比较分析。基本算法如下:

$$M(x,y) = \log_2\frac{p(x,y)}{p(x)p(y)}$$ 公式(12)[20]

其中,$p(x,y)$表示某两分词的联合概率分布,即词条 x 与 y 在用户生成内容中共同出现的概率,$p(x)$与 $p(y)$则对应词条 x 与词条 y 分别在用户生成内容中的概率分布。一般而言,$M(x,y)$越大,则说明他们两者之间的关系越紧密,可能同时出现的几率就越大,反之则说明共现几率越小。

4 研究结果

4.1 实验数据

医享网是国内在线健康社区中用户较多,可信度较高的社区之一,支持病例库查询,在线健康问题问答,其中痛风圈社区的内容交互较为频繁,论述相对全面[21]。故本研究设置医享网的痛风圈作为数据来源,收集时间为 2020 年 2 月,依据相关公开内容,设计爬虫程序进行数据抽取,具体数据主要包括用户昵称、发帖与回帖内容。

进一步对数据进行预处理,分词处理,即使用 JIBEA 对用户文本进行分词,以进行高频词统计和互信息模型构建;用户编码映射表构建,即对所有用户进行统一编码,如用户 1、用户 2 等顺排至最后;用户评论映射表,即对用户所发表的评论内容进行对应;用户评论者映射表,即构建评论用户 1、用户 2 等的用户评论映射表;异常用户处理即过滤掉评论或发帖与通风圈无关的用户,如推送广告用户等。经过最终预处理,共得到 292 位有效用户的 2 560 条有效交互内容。

4.2 分析结果

4.2.1 基于 PageRank 的用户交互影响力分析结果

用户交互影响力的计算主要通过本研究公式(4)中的 PageRank 的算法进行计算,具体分析结果见图 2。从整体分布来看,大多数用户的交互影响力偏低,而只有少部分用户处于高影响力水平,如用户 253、用户 269、用户 151、用户 154 等,说明这部分用户受到其他用户的较多关注,具有一定的影响力。

但基于 PageRank 算法的排名,仅考虑了用户之间的交互机制来加以判定,虽然具有一定的实用性,但忽略了信息使用价值的判断,即一些用户尽管交互影响力很高,但其所发布的信息如果受到其他多数用户的质疑或者否定,那该信息的有效性会受到影响,如评论用户对高交互影响力用户 253 和用户 269 的评价是分别存在“你是中医?”“就是因为无法治愈”“是不是庸医忽悠你呢”“是你自己无知”等负面或质疑倾向的评论内容,这会在一定程度上影响该用户的交互影响力。因此,本研究将进一步探讨评论情感倾向性对用户影响力的影响,并探索将评论情感倾向融合进入用户的交互影响力值中,综合探讨和分析用户生成内容的使用价值,从而提升用户影响力测量的客观性和有效性。

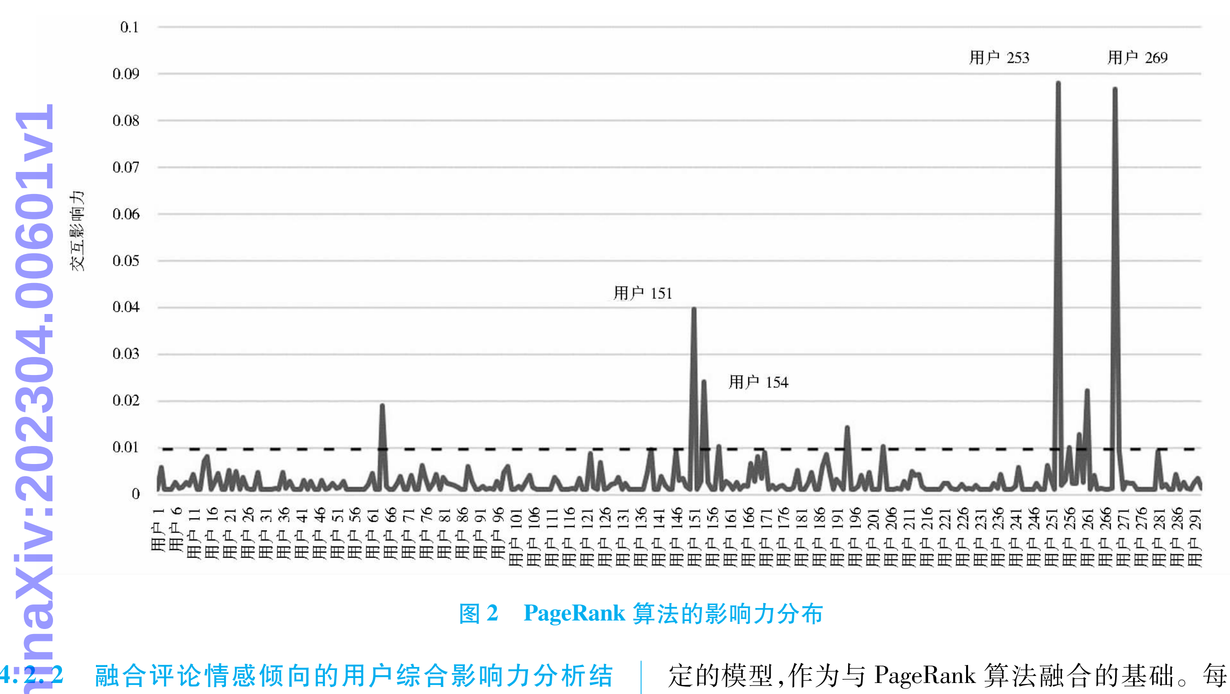


图 2 PageRank 算法的影响力分布

4.2.2 融合评论情感倾向的用户综合影响力分析结果

(1)情感倾向分类模型的选择与分析。本研究在 PageRank 分析结果的基础上,融合用户情感倾向的分析,对相关文本进行分析。为确定评论情感倾向分析的最优模型,本文选取了机器学习算法中的 RandomForest(随机森林算法)、Logistic(逻辑回归算法)、SGD(随机梯度下降算法)、SVM(支持向量机算法)以及 Bayesian(朴素贝叶斯算法)五大经典算法进行比较分析。首先对文本中的情感倾向进行两轮的人工数据标注,所判断的一致性达到 95% 以上。同时为优化情感倾向识别效果,通过多轮试验与调试,最终确定主要参数设置值:随机森设置林 min_samples_leaf 为 1,min_samples_split 为 2,criterion 为“gini”算法,n_estimator 为 10;SGD 设置 loss 为“log”,max_iter 为 100;SVM 设置 kernal 为“linear”,C 为 1;Logistic 与朴素贝叶斯均采用默认参数进行比较判断,从而选择综合性能较高,更稳定的模型,作为与 PageRank 算法融合的基础。每次训练都重新对训练集数据进行评估,测试集设占总数据量的 20%,训练集占 80%,分别迭代 10 次,具体计算结果见表 1,可以发现,Logistic 回归算法 F1 值相对较低,说明模型效果一般,而基于线性函数的 SVM 模型 10 次 F1 平均值(AVEG_F1)与平均准确度(AVEG_ACC)都是最高的,略优于其他算法,其方差最小(S2_F1),具有更加稳定的预测能力,故选 SVM 模型对整体数据进行识别和分类。

其次,在确定采用 SVM 模型进行计算的基础上,对所有用户对应的评论情感倾向性进行了分析和对比,具体计算结果见表 2 和图 3,其中用户 1 的交互影响力为 0.001,情感倾向值为 0,用户 7 的交互影响力为 0.001,情感倾向值为 0.071,而用户 151,用户154 等用户有较高的情感倾向值,分别是 1.797 与 1.294,但其交互影响力较低,只有 0.040 与 0.024;而用户 253 和用户 269 则具有较低的情感倾向值,分别是0.142与 0.071,但有

表 1　机器学习各个模型的计算结果比较

第 i 次	RandomForest	Logistic	SGD	SVM	Bayesian
1	0.85	0.86	0.86	0.88	0.86
2	0.94	0.86	0.87	0.92	0.90
3	0.88	0.88	0.90	0.86	0.89
4	0.86	0.84	0.91	0.90	0.88
5	0.88	0.86	0.84	0.89	0.92
6	0.92	0.83	0.88	0.91	0.93
7	0.88	0.82	0.93	0.95	0.91
8	0.80	0.88	0.85	0.91	0.86
9	0.86	0.80	0.94	0.91	0.86
10	0.92	0.89	0.91	0.93	0.90
AVEG_F1	0.88	0.85	0.89	0.91	0.89
AVEG_ACC	0.87	0.89	0.91	0.92	0.88
S^2_ F1	0.002	0.001	0.001	0.001	0.001

较大的交互影响力，达到了0.088与0.087。

最后，从整体数据来看，用户的评论情感倾向的区分度较为明显，与所对应用户的交互影响力们的分布不完全一致，可以作为综合影响力重要指标之一，对交互影响力进行融合和补充。

表 2　用户交互影响力和评论情感倾向分布(随机部分)

用户	交互影响力	评论情感倾向值	用户	交互影响力	评论情感倾向值
1	0.001	0.000	113	0.003	0.247
7	0.001	0.071	117	0.001	0.071
13	0.001	0.000	138	0.005	0.734
17	0.003	0.106	139	0.010	0.667
32	0.001	0.000	151	0.040	1.797
42	0.003	0.177	154	0.024	1.295
48	0.001	0.036	175	0.002	0.901
68	0.002	0.036	253	0.088	0.142
74	0.001	0.000	256	0.010	0.383
94	0.002	0.035	269	0.087	0.071

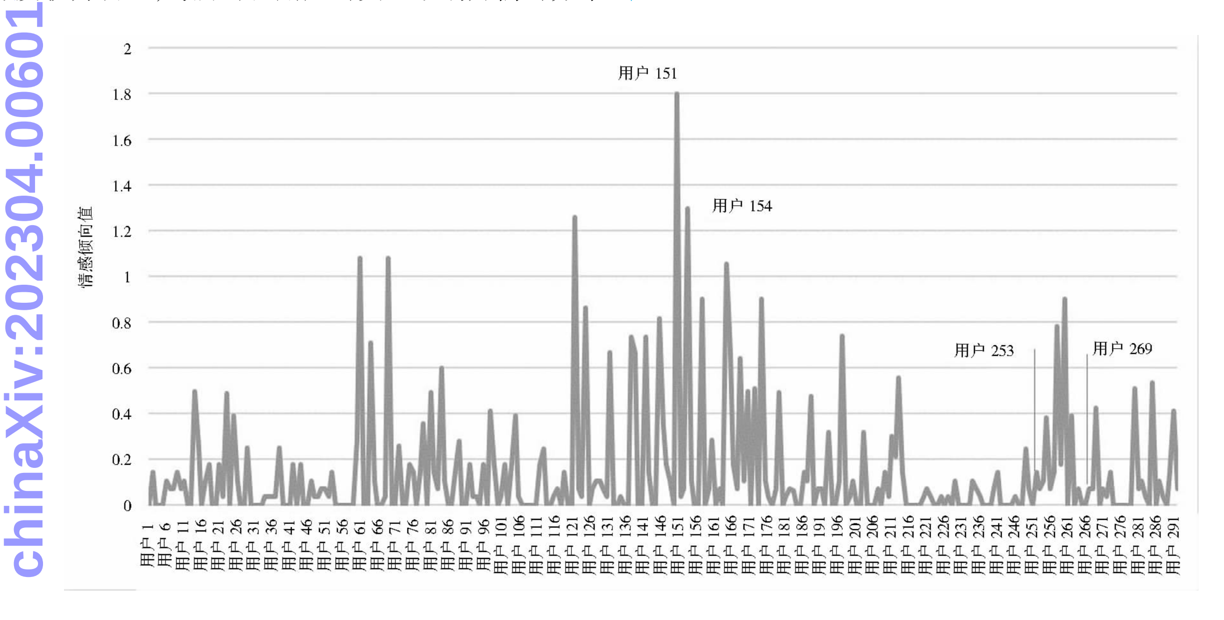


图 3　用户评论情感倾向值分布

(2)用户综合影响力分析结果。在以上研究的基础上，进一步将情感倾向和交互影响力进行融合分析，并得出综合影响力分布，其结果见表 3 和图 4。图 4 中，横坐标代表的是 292 位用户，纵坐标代表的是每位用户融合情感倾向值与用户交互影响力 PR 值后的结果，即综合影响力。大多数用户在 0 到 0.01 范围内，有部分用户数值相对较大，在 0.01 水平以上，最高达到了 0.07 水平左右。其中，用户 151 综合影响力得到了较大的提升，达到了 0.071 5，有 29% 的其他用户对其内容具有较强的正向情感倾向，这对于用户的综合影响力具有较大的影响；用户 154 达到了 0.031 2，有 33% 的用户持有积极情感倾向，但同时有 13% 的用户具有消极情感倾向。此外，通过比较发现发现，用户 253 和用户 269 等人的综合影响力相较于自身的交互影响力变化也较大，见图 4。对比图 2 发现，用户交互影响力排名较高，但如果内容并不能得到多数评论的肯定，其综合影响力排名大多会有较大变化。这也说明在线健康社区中，并非所有的用户或者内容都是符合用户需求的，有些用户虽然交互影响力较高，但其可能是处于寻求帮助的状态，甚至部分用户所提到的内容可能具有广告性质或灌水行为，并未得到相应的肯定评论，甚至得到较多的负面评价，因此其不应作为有价值的信息，发布信息的用户的综合影响力也将受到相应的影响。

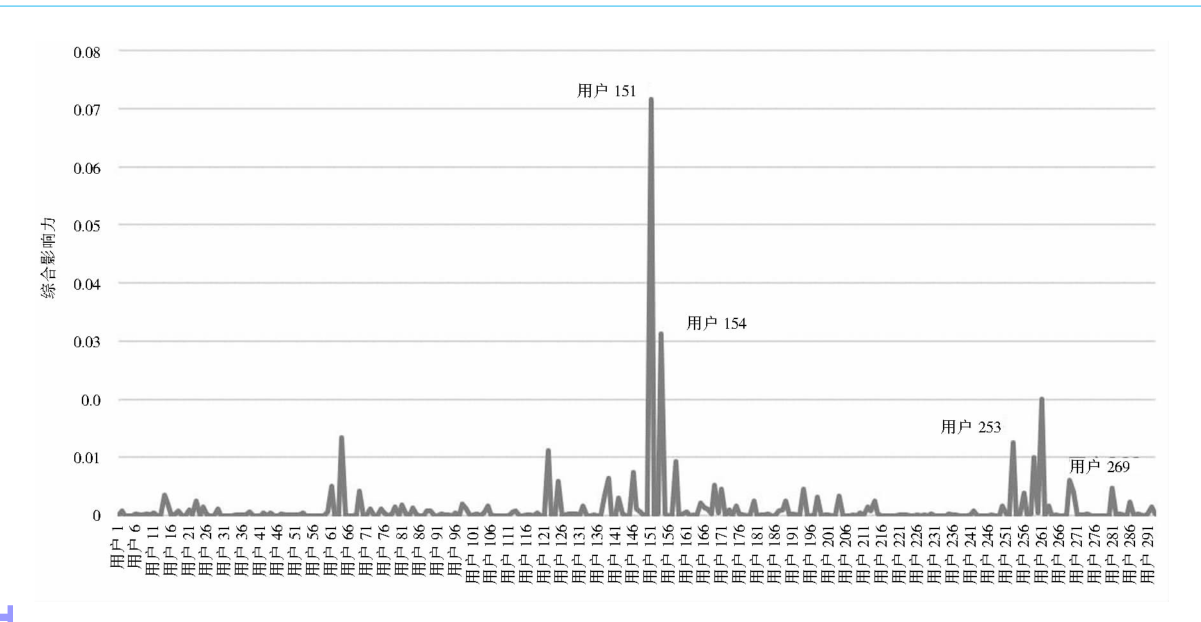


图 4 综合影响力分布

表 3 综合影响力排名结果(随机部分)

用户	综合影响力	用户	综合影响力
1	0.000 0	113	0.000 7
7	0.000 1	117	0.000 1
13	0.000 0	138	0.003 5
17	0.000 3	139	0.006 4
32	0.000 0	151	0.071 5
42	0.000 5	154	0.031 3
48	0.000 0	175	0.001 6
68	0.000 1	253	0.012 5
74	0.000 0	256	0.003 9
94	0.000 1	269	0.006 2

另外,为进一步探究不同用户交互影响力到综合影响力的转变机理,本研究对具体用户的各项指标与评论数据内容进行了梳理,以上述图中所标注的 4 位典型用户(用户 151、用户 154、用户 253、用户 269 等)作为案例对象,具体结果见表 4。由于用户 151、用户 154 收到的评论数据多为"学习了,谢谢""顶一下""感谢感谢"等积极类文本,因此这类用户情感倾向值较大,具有较重要的实用价值和传播意义,在整体上提高了用户的综合影响力。而用户 253 和用户 269 虽具有较高的交互影响力,但由于其得到的评论情感多为质疑和消极,如"你是中医?""真的可以治愈吗?""是你自己无知",因此评论情感倾向值较低,从而使得其综合影响力受到影响而下降。

表 4 典型用户实验对比

影响力转变用户	低交互转高综合		高交互转低综合	
	用户 151	用户 154	用户 253	用户 269
交互影响力	0.040	0.024	0.088	0.087
情感倾向值	1.797	1.295	0.142	0.071
综合影响力	0.072	0.031	0.013	0.006
评论数据	学习了,谢谢	感谢分享偏方	真的可以治愈吗??	术前照片有吗?
核心观点	说的有道理,看到了新思路	这个一定要顶!!	你是中医?	是不是庸医忽悠你呢
	受教了,多谢	新方法	现在医学有这个技术吗?	不搞微创手术? 直接一大刀。吓人啊
	有点意思	感谢感谢!	希望能够造福世人	是你自己无知
	小手一抖,经验到手	明天就试试看	听说西药吃了不好又没怎么吃	没必要开这么大条口吧?
	顶一下	谢谢楼主分享	就是因为无法治愈	不是吧? 一年发展成这样
	科普贴,支持	这贴得顶啊!	可以喝百草清风茶,很有效果	严重怀疑,不能称痛风石
	说的很对	收藏	多吃蔬菜少吃海鲜	祝福、加油

4.3.3 基于互信息的综合影响力用户文本内容分析结果

为进一步探究高综合影响力用户对健康社区主题方向的影响，本研究选取了前 20 个高综合影响力用户和社区其他用户评论内容进行实验，通过 TF-IDF 与互信息两种算法，构建了该用户群的词条共现网络，为了更直观和清楚展示其关联效果与整体结构，借助 Vosviewer 软件对该共现网络进行了可视化分析，具体结果见图 5。其中，疼痛、结晶、血症、高尿酸、发作、关节、止痛、代谢、含量等半径较大的 15 个节点代表高频词汇，而半径较小的其他节点分别表示每个高频词的 10 个互信息关键词，相连表示对应的一组数据。由图 5 可以得知，高综合影响力用户主要关注的几个问题集中在三个方面，其一是当痛风病发作的时候的表现，如疼痛、临睡前、侵蚀、结晶、沉淀、磷脂、关节、神经等关键词均在不同方面说明了病症发生的症状、时间等内容；其二是治疗痛风所使用的药物，包括秋水仙碱、双氯芬酸钠、阿西美辛、戴芬、秋水仙、别嘌呤、抗炎药、止痛针等内容；其三，为了更好的治疗痛风所采用的食疗辅助，如特别注意龙须菜、菠菜、蘑菇、鲜豌豆类含嘌呤较多的蔬菜，蛤类、动物内脏、多饮水、少喝汤等。

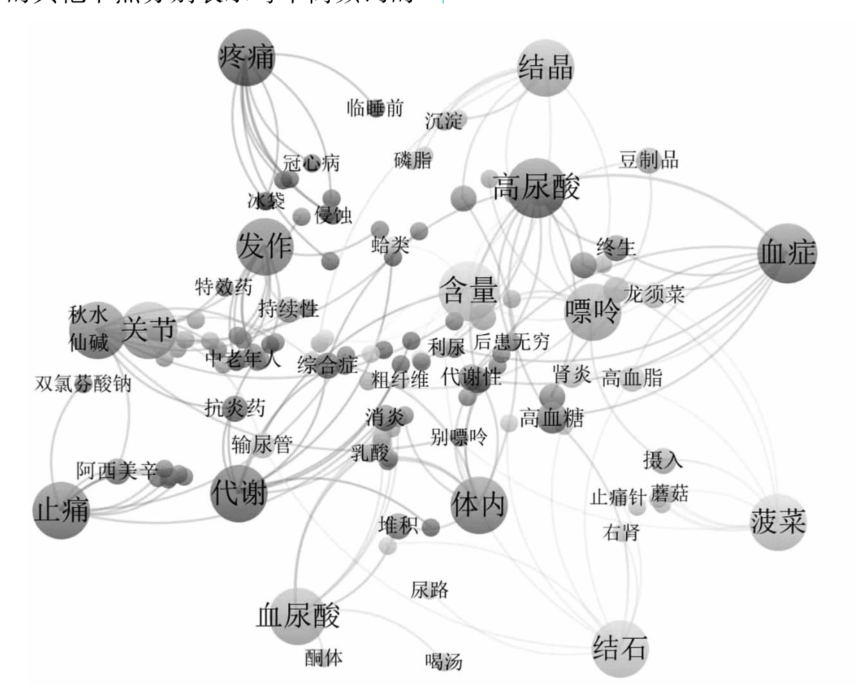


图 5　高综合影响力用户发布信息共现可视化结果

同理，我们将社区中其他用户生成的信息高频词与互信息共现矩阵进行了分析，其具体结果见图 6。其中，大致也能够分为三个主要方面，其一是疾病的具体症状，糜烂、肿痛、剧痛、挛缩、发麻、发红等关键词；其二是药物治疗，如秋水仙碱、阿西美辛、别嘌醇、双氢克尿塞、次黄嘌呤、英太青、布洛芬、散热等药物和方法；其三是辅助治疗，通过关键词可以发现，一方面是对豆皮、火腿肠、荤菜等食品尽量避免，同时可以对蔬菜水果、茭白、无机盐以及相关的偏碱性等食物进行摄入。

总体来看，高综合影响力用户所关注的内容主题与社区中其他用户信息所关注的主题具有较强的一致趋向性，即高综合影响力用户与多数用户讨论的核心内容一致。这在一定程度上说明本研究中所识别出的高影响力用户所发布的信息具有较强的使用价值，他们所发布的主题内容在一定程度上引导了该社区中的信息发布方向，也在一定程度上说明所构建的综合影响力的识别和分析方法具有一定的合理性和客观性。同时，对于健康信息的需求者来说，有效的识别出高影响力用户及其相关的信息，可以更好的节约需求者的信息查询成本，快速的了解到本社区的特点和相应的主题方向，有助于提升需求者的健康行为决策的效率和效果。

5　总结与展望

5.1　研究总结

本研究从交互影响力和情感倾向两个视角出发，基于 SVM 算法的线性核函数建立了情绪识别模型，对

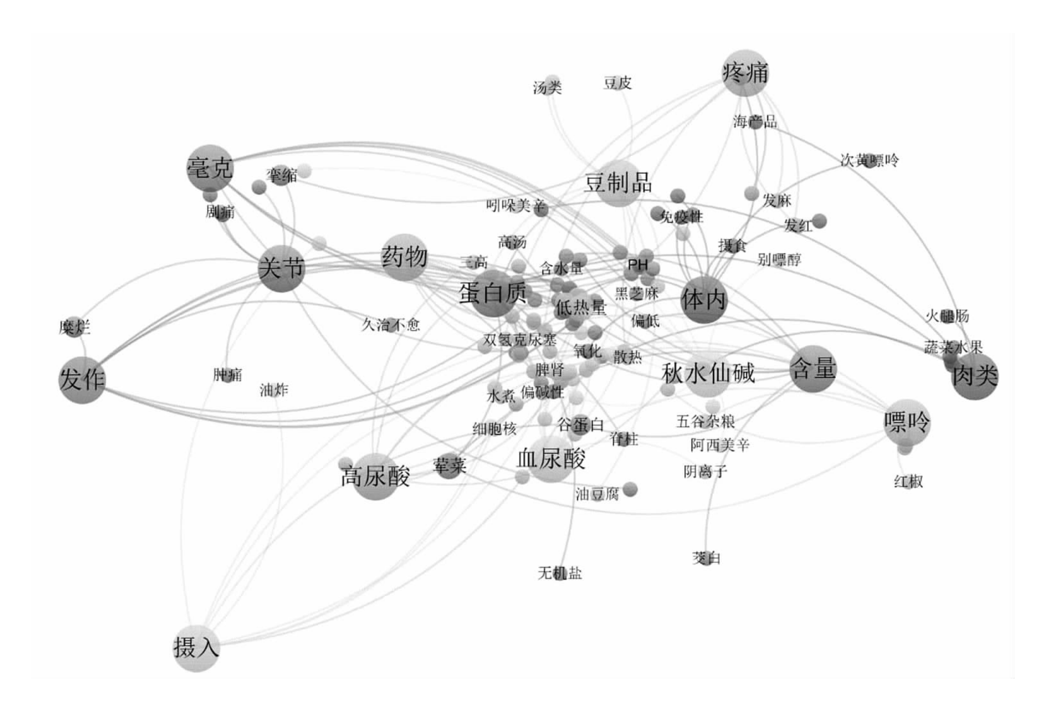


图6　社区其他用户发布信息共现可视化结果

文本有效性进行了识别与分析，并结合交互影响力与情感倾向性探讨了用户综合影响力，得出以下结论：

首先，本研究通过对用户交互影响力的计算，并结合相关案例，发现交互影响力更多强调了交互活跃性，但对于揭示用户信息资源的有效性方面存在一定不足，因此并不能完全客观的反映出用户的真实影响力，需要引入评论情感倾向值对影响力做进一步的融合计算。

其次，通过对比机器学习的五种关于情感计算的主要算法中，发现SVM算法对于本文构建的评论情感倾向分类模型效果最优，为有效的计算用户综合影响力提供了技术支撑。

再次，本研究探索性地将PageRank交互影响力与评论情感倾向进行融合计算，并通过相应个案分析从信息内容的角度对高影响力用户做了进一步验证，在一定程度上说明了本研究中的综合影响力的算法具有较好的合理性和适用性。

此外，通过对高综合影响力的高频词-互信息矩阵与其他用户群体的高频词-互信息矩阵的比较发现，二者相似程度较高，基本主题方向一致，这也在一定程度上说明了寻找高综合影响力用户的必要性，也进一步说明本研究的用户影响力综合计算方法可较为客观的识别出主导健康社区内容方向的具有较高影响力的用户，有助于健康信息需求者能够及时、准确从健康社区中获得所需有价值的信息，提升健康信息的利用效果。

5.2　研究展望

本文提出了一种情绪识别模型以探索用户生成内容的情感倾向，从而构建用户综合影响力的研究方法，并进一步通过具体内容分析论述了高综合影响力用户对社区方向的影响，但也存在一定的不足：

（1）交互影响力和情感分析算法的优化。本研究用户交互影响力主要基于PageRank算法，虽然该方法应用较为广泛，但其在分析用户影响力方面仍存在一定的改进空间，可在今后的研究中，结合用户行为特征对该算法做进一步的优化。此外，本研究中所用到的情感倾向分析的算法，今后可对更多的相关算法和框架进行比较分析，从而进一步提升相关计算的效率和准确性。

（2）研究数据的进一步丰富。本研究主要基于医享网的痛风病圈数据进行了研究，在今后的研究中，可以进一步扩展健康社区的数据获取范围，通过比较不同健康社区中用户综合影响力的分布和特征，以不断拓展和验证本研究的适用性。

参考文献：

[1] 杨梓. 最新！卫健委发布全国医疗相关数据[EB/OL]. [2021-04-27]. https://www.sohu.com/a/247593213_439958.

[2] 2018年中国健康医疗大数据行业发展现状及发展趋势分析[EB/OL]. [2021-04-27]. http://www.chyxx.com/industry/201806/649591.html.

[3] 国务院办公厅. 国务院办公厅关于促进“互联网+医疗健康”发展的意见[EB/OL]. [2021-04-27]. http://www.pkulaw.cn/fulltext_form.aspx? Db=chl&Gid=37395b41f6f018e4bdfb&keyword=%E5%

8C% BB% E7% 96% 97&EncodingName = &Search _ Mode = accurate&Search_IsTitle =0.
[4] 王佳敏,吴鹏,陈芬,等. 突发事件中意见领袖的识别和影响力实证研究[J]. 情报学报, 2016, 35(2):169－176.
[5] 赵发珍. 基于链接分析法的网络社区影响力研究——以国内30个网络社区网站为例[J]. 现代情报,2013,33(6):91－95.
[6] 董伟,李建红,陶金虎. 在线健康社区活跃用户识别及其交互类型分析[J]. 文献与数据学报,2020,2(1):89－101.
[7] 陈远, 刘欣宇. 基于社会网络分析的意见领袖识别研究[J]. 情报科学, 2015, 33(4): 13－19, 92.
[8] 谢英香,冯锐. 虚拟教师社区中博客网络位置的影响力研究[J]. 现代教育技术,2010,20(1):97－100,110.
[9] JONNALAGADDA S, PEELER R, TOPHAM P. Discovering opinion leaders for medical topics using news articles[J]. Journal of biomedical semantics,2012,3(1):2.
[10] 陈芬,高小欢,彭玥,等. 融合文本倾向性分析的微博意见领袖识别[J]. 数据分析与知识发现,2019,3(11):120－128.
[11] 刘玲,杨长春. 一种新的微博社区用户影响力评估算法[J]. 计算机应用与软件,2017,34(7):212－216,261.
[12] 张俊豪,顾益军,张士豪. 基于 PageRank 和用户行为的微博用户影响力评估[J]. 信息网络安全,2015(6):73－78.
[13] SONG X, CHI Y, HINO K, et al. Identifying opinion leaders in the blogosphere[C]// ACM. Proceedings of the sixteenth ACM conference on conference on information and knowledge management. Lisbon: ACM, 2007:971－974.
[14] 韩忠明,苑丽玲,杨伟杰,等. 加权社会网络中重要节点发现算法[J]. 计算机应用,2013,33(6):1553－1557,1562.
[15] 肖宇, 许炜, 商召玺. 微博用户区域影响力识别算法及分析[J]. 计算机科学, 2012, 39(9): 38－42.
[16] 马凤. 基于 PageRank 算法的期刊影响力研究[J]. 情报杂志, 2014,33(12):103－108.
[17] ZHANG Y. Determinants of poster reputation on internet stock message boards[J]. American journal of economics and business administration,2009,1(2):114.
[18] Precision,Recall,F1score,Accuracy 的理解[EB/OL].[2021－04－27]. https://blog. csdn. net/u014380165/article/details/77493978.
[19] ALAM S, YAO N. Big data analytics, text mining and modern english language[J]. Journal of grid computing, 2019,17(2): 357－366.
[20] 费洪晓,康松林,朱小娟,等. 基于词频统计的中文分词的研究[J]. 计算机工程与应用,2005(7):67－68,100.
[21] 董伟,陶金虎. 基于主题偏好的在线健康社区用户兴趣群体识别研究——以医享网为例[J]. 情报科学,2021,39(3):88－93,119.

作者贡献说明：

董伟：撰写论文，修改论文；
陶金虎：数据爬取与分析，撰写论文。

Research on the User's Influence in Online Health Community Based on PageRank and Emotional Tendency

Dong Wei Tao Jinhu
School of Education, Tianjin University, Tianjin 300350

Abstract: [**Purpose/significance**] The effective identification of high-impact users in online health communities is helpful for demanders to find valuable health information, which is of great significance for reducing the cost of health information search and improving the effectiveness of health behavior decision-making. [**Method/process**] This study was from the perspective of interactivity of users and emotional tendency of comments, using PageRank and SVM algorithm to build a method to measure the users' influence in online health community, and took the medical network as experimental object, from the angle of content use value, further calculated the comprehensive influence of users in the community, and in case the user is analyzed. [**Result/conclusion**] The results show that the algorithm is reasonable and can optimize the calculation results of PageRank algorithm. At the same time, the TF-IDF and Mutual Information algorithm are used to reveal that the information content published by high comprehensive influence users is basically consistent with content topics of other user groups in the community, and such users play a certain role in guiding the theme direction of the community. Therefore, the method constructed in this study can effectively and reasonably identify high-impact users, which is helpful for health demanders to find the required information timely and accurately, improving the effect of using health information, so as to enrich the theoretical and practical research on the information behavior of users in online health communities.

Keywords: PageRank emotional tendency online health community user influence